

Temas actuales: Next Generation Sequencing (NGS) y secuenciadores personales

Bioinformática

22-3-19

Elvira Mayordomo

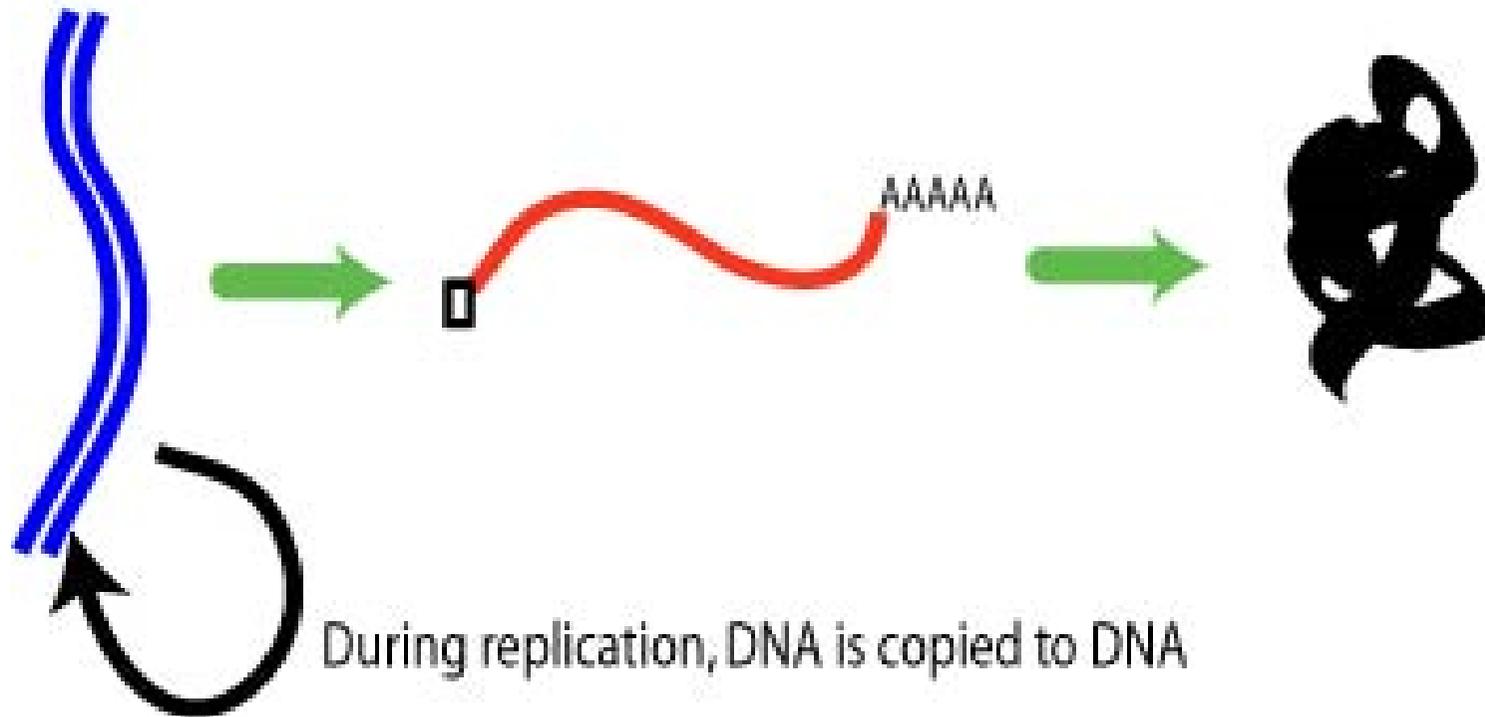


Veremos

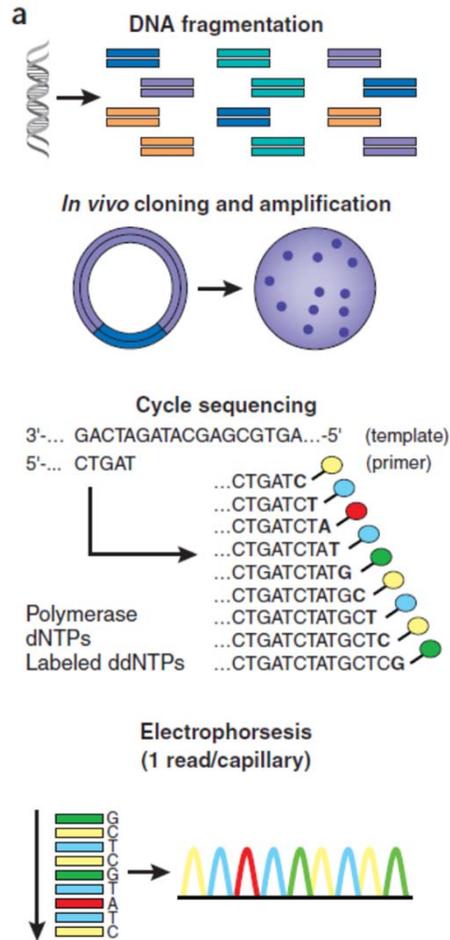
- **Historia**
- Plataformas NGS
- Tercera generación: Secuenciadores personales
- Aplicaciones
- Retos bioinformáticos

Dogma central

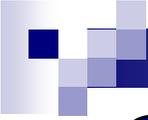
DNA is transcribed to RNA is translated to PROTEIN



Secuenciación de Sanger



- El DNA se fragmenta
- Clonación
- Reacción de secuenciación cíclica
- Separación por electroforesis
- Lectura con etiquetas fluorescentes



Sanger vs NGS

- La secuenciación de Sanger ha sido el **único método de secuenciación** de DNA durante **30 años**, pero ...
- ... había gran **necesidad** de una **mayor eficiencia** y de tecnologías de secuenciación **más económicas** ...
- NGS tiene la capacidad de **procesar en paralelo** millones de lecturas de secuencia en lugar de 96 a la vez (1/6 del coste)
- **Objeciones**: la fidelidad, la longitud de lectura, el coste de infraestructura, el manejo de grandes volúmenes de datos



NGS

- NGS = Secuenciación amplificada de una única secuencia
- Third generation sequencing = secuenciación de una única molécula



Veremos

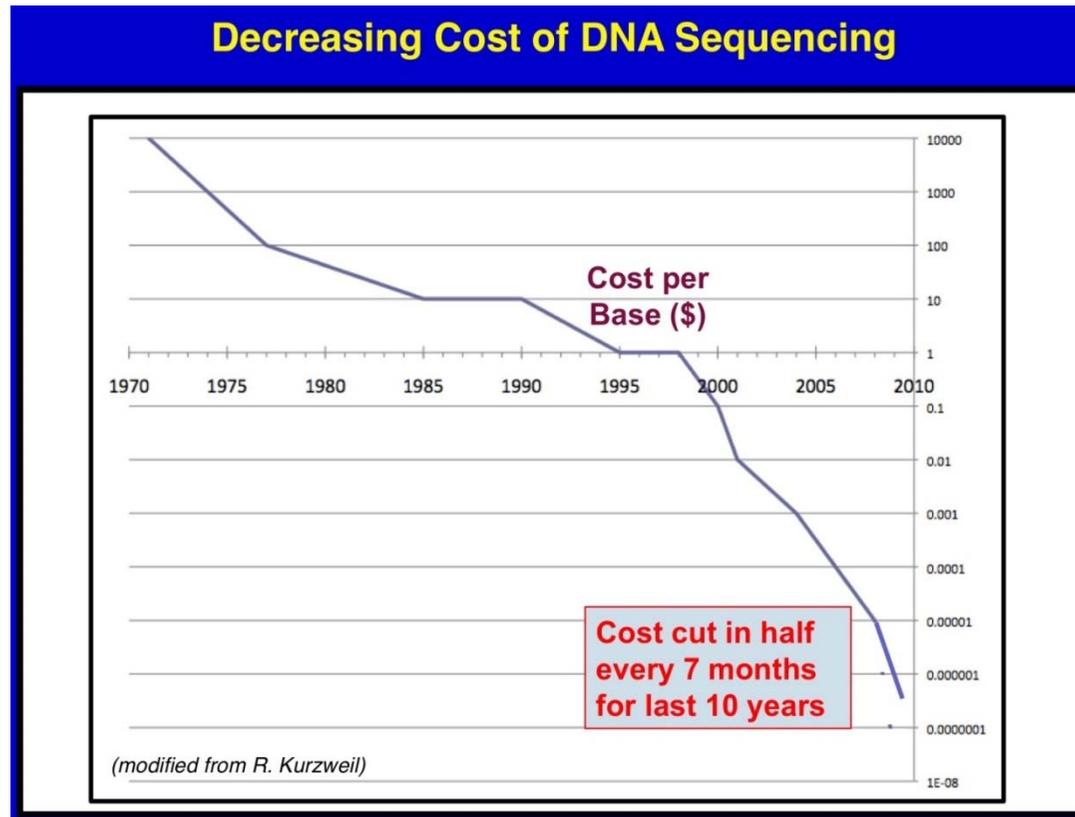
- Historia
- **Plataformas NGS**
- Tercera generación: Secuenciadores personales
- Aplicaciones
- Retos bioinformáticos

Plataformas NGS

- Roche / 454 FLX: 2004
- Illumina Solexa Genoma Analyzer: 2006
- Applied Biosystems SOLiD™ System: 2007
- Helicos Heliscope™ : 2008
- Pacific Biosciences SMRT: 2010



Reducción de costes





Tres plataformas de secuenciación principales

- Roche 454
- **Illumina Solexa**
- Applied Biosystems SOLiD



Esquema general

1. **Preparación biblioteca**
2. 2 opciones
 1. PCR por emulsión
 2. PCR polony en una transparencia
3. 3 opciones
 1. Amplificación de puentes (Illumina)
 2. Amplificación por temperatura (SOLiD)
 3. Pirosecuenciación (454)



1 Preparación de la biblioteca

- Se añaden a los fragmentos
 - Adaptadores
 - Primers (principios para replicación)
 - Códigos de barras

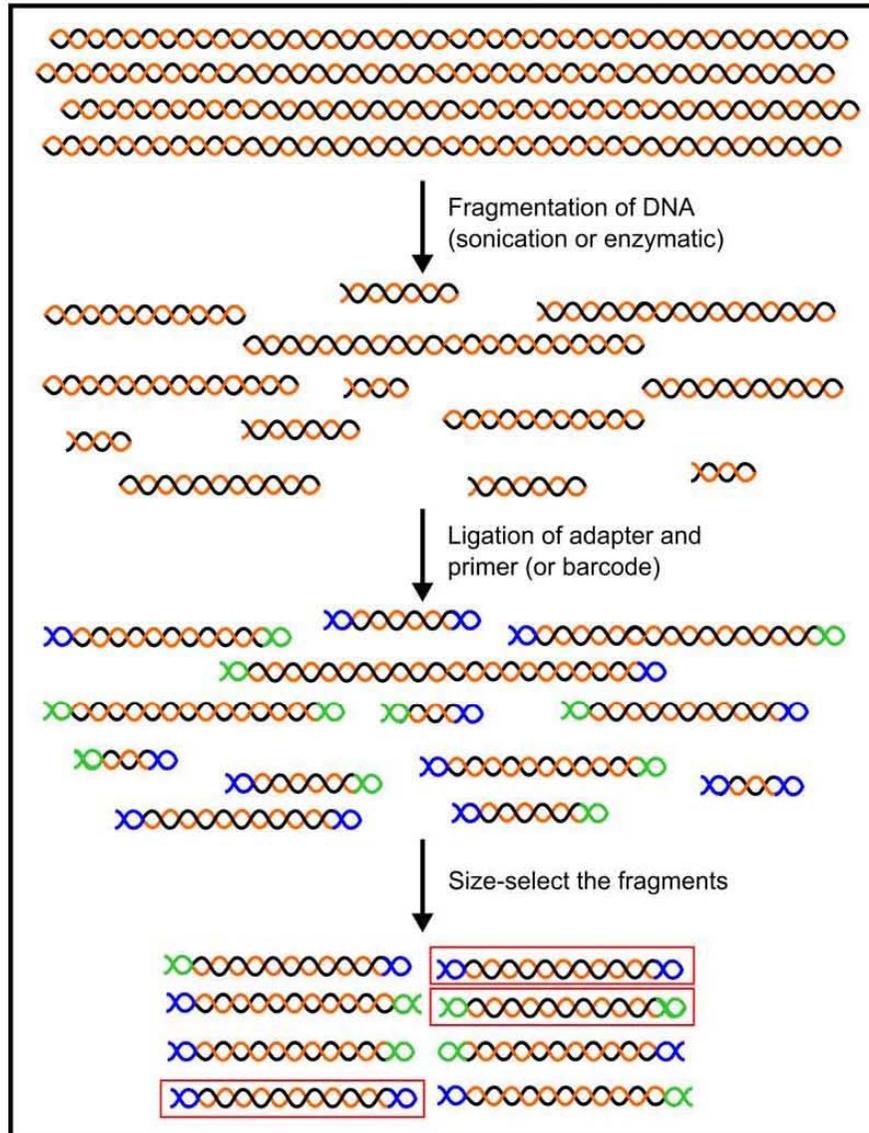
Next Generation Sequencing Workflow

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

CTAGGTAGCTAGTCTG
GCTLIFECISGATAG
C4-LETTERWORDT
GCTATATCGTAGCTG

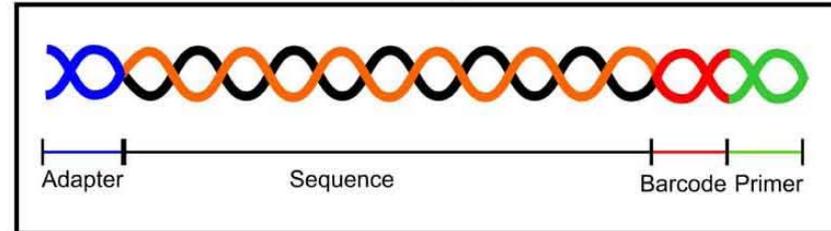


Next Generation Sequencing : Amplified Single Molecule Sequencing



Library preparation

Good fragments :



1- Preparación biblioteca



Esquema general

1. Preparación biblioteca
2. 2 opciones
 1. **PCR por emulsión**
 2. PCR polony en una transparencia
3. 3 opciones
 1. Amplificación de puentes (Illumina)
 2. Amplificación por temperatura (SOLiD)
 3. Pirosecuenciación (454)



2.1 PCR por emulsión

- Se añaden microesferas recubiertas con cebadores (beads)
- El fragmento de DNA se desnaturaliza en 2 hebras
- Una hebra se une a la microesfera y empieza a replicarse
- Se consiguen de 30 a 60 copias

Next Generation Sequencing Workflow

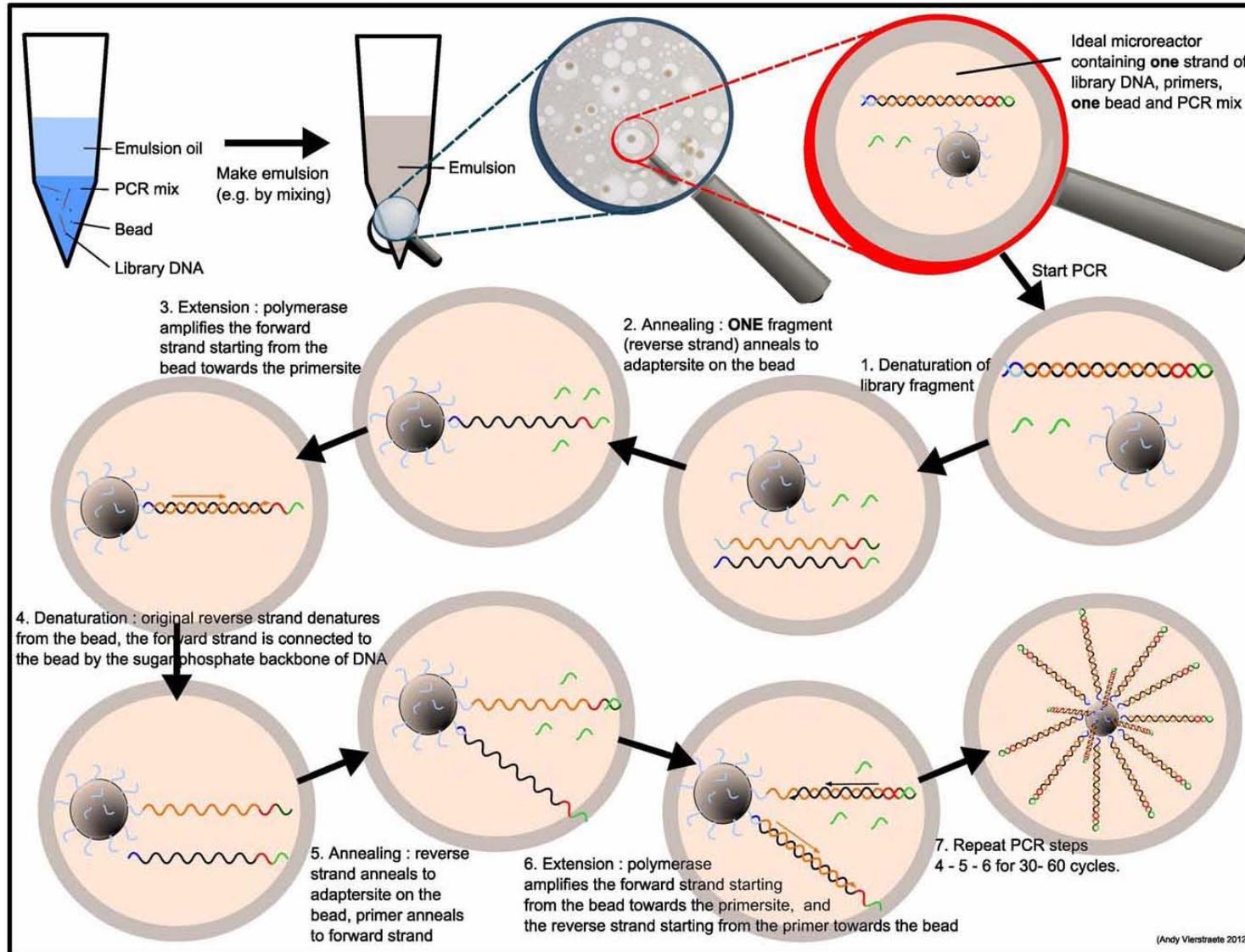
Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

CTAGGTAGCTAGTCC
GCTLIFECISGATAG
C4-LETTERWORDT
GCTATATCGTAGTG



Next Generation Sequencing : Amplified Single Molecule Sequencing Emulsion PCR

11/132





Esquema general

1. Preparación biblioteca
2. 2 opciones
 1. PCR por emulsión
 2. **PCR polony en una transparencia**
3. 3 opciones
 1. Amplificación de puentes (Illumina)
 2. Amplificación por temperatura (SOLiD)
 3. Pirosecuenciación (454)



2.2 PCR polony en una transp.

- Pegar el DNA a una superficie con muchos primers
- Provocar la formación de puentes

Next Generation Sequencing Workflow

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

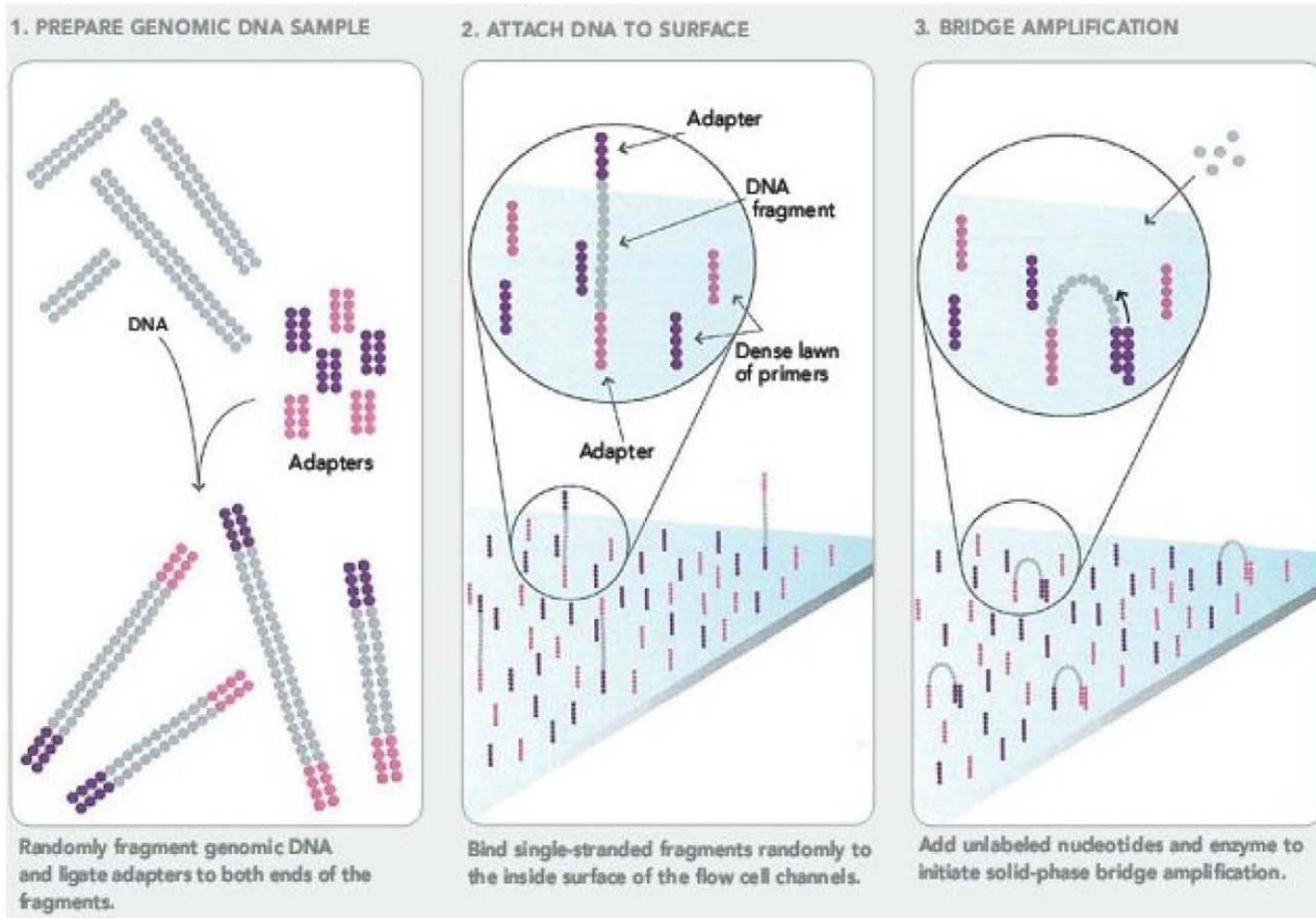
CTAGGTAGCTAGTCC
GCTLIFECISGATAG
C4-LETTERWORDT
GCTATATCGTAGCTG



Next Generation Sequencing : Amplified Single Molecule Sequencing “Polony” PCR

18/132

Bridge amplification : Illumina





Esquema general

1. Preparación biblioteca
2. 2 opciones
 1. PCR por emulsión
 2. PCR polony en una transparencia
3. 3 opciones
 1. **Amplificación de puentes (Illumina)**
 2. Amplificación por temperatura (SOLiD)
 3. Pirosecuenciación (454)



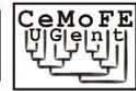
3.1 Amplificación de puentes

- Se construyen dobles puentes
- Se rompen los puentes (el resultado es duplicación)
- Repetir el proceso construcción puentes – duplicación - rotura

Next Generation Sequencing Workflow

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

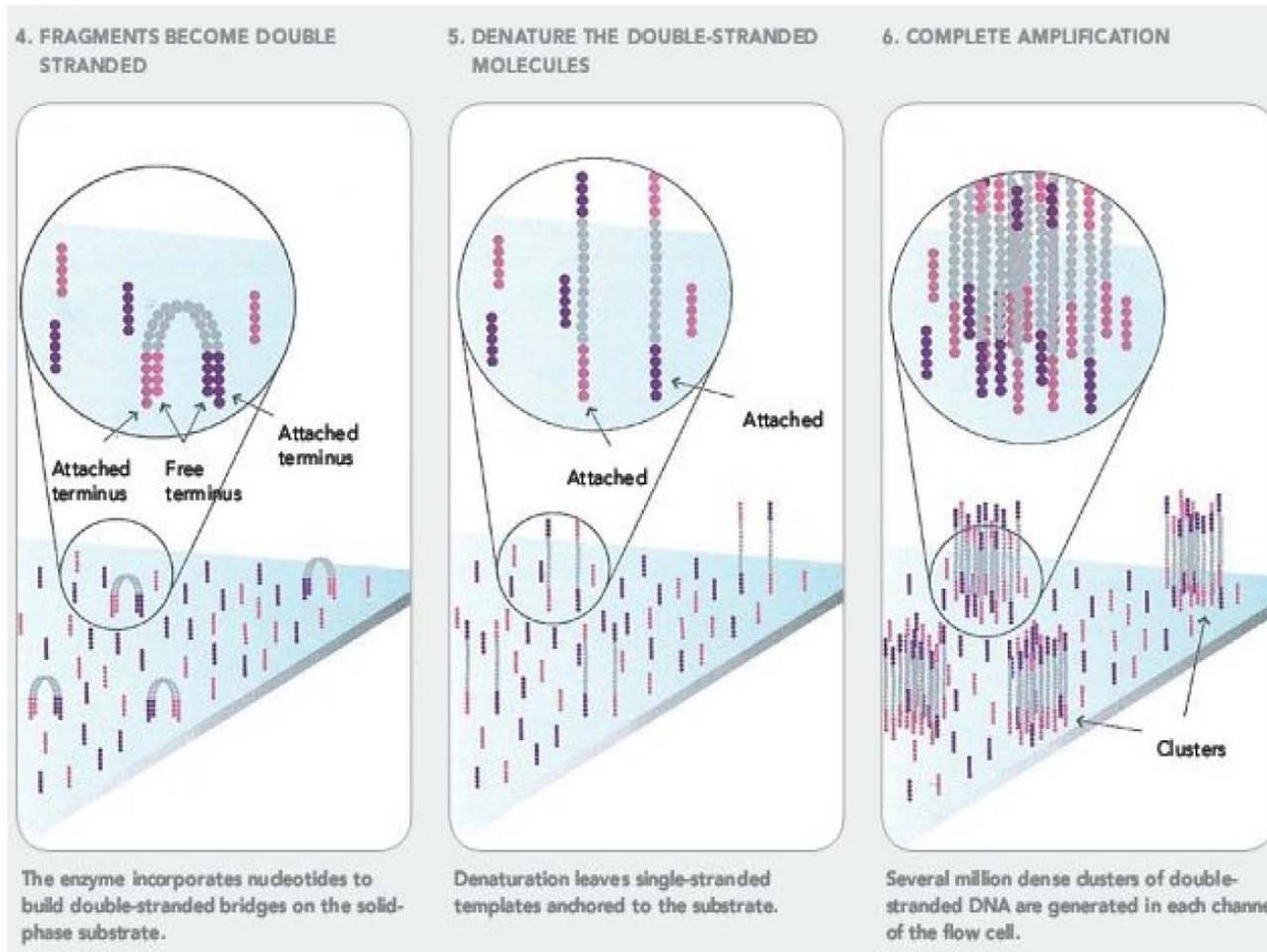
CTAGG TAGCTAGTCTG
GCTLIFECISGATAG
C4-LETTERWORDT
GCTATATCGTAGCTG



Next Generation Sequencing : Amplified Single Molecule Sequencing “Polony” PCR

Bridge amplification : Illumina

19/132





Esquema general

1. Preparación biblioteca
2. 2 opciones
 1. PCR por emulsión
 2. PCR polony en una transparencia
3. 3 opciones
 1. Amplificación de puentes (Illumina)
 2. **Amplificación por temperatura (SOLiD)**
 3. Pirosecuenciación (454)



3.2 Amplificación por temperatura

- Replicación y desplazamiento (andar) con distintos cambios de temperatura
- Todo a través de los enzimas adecuados

Next Generation Sequencing Workflow

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

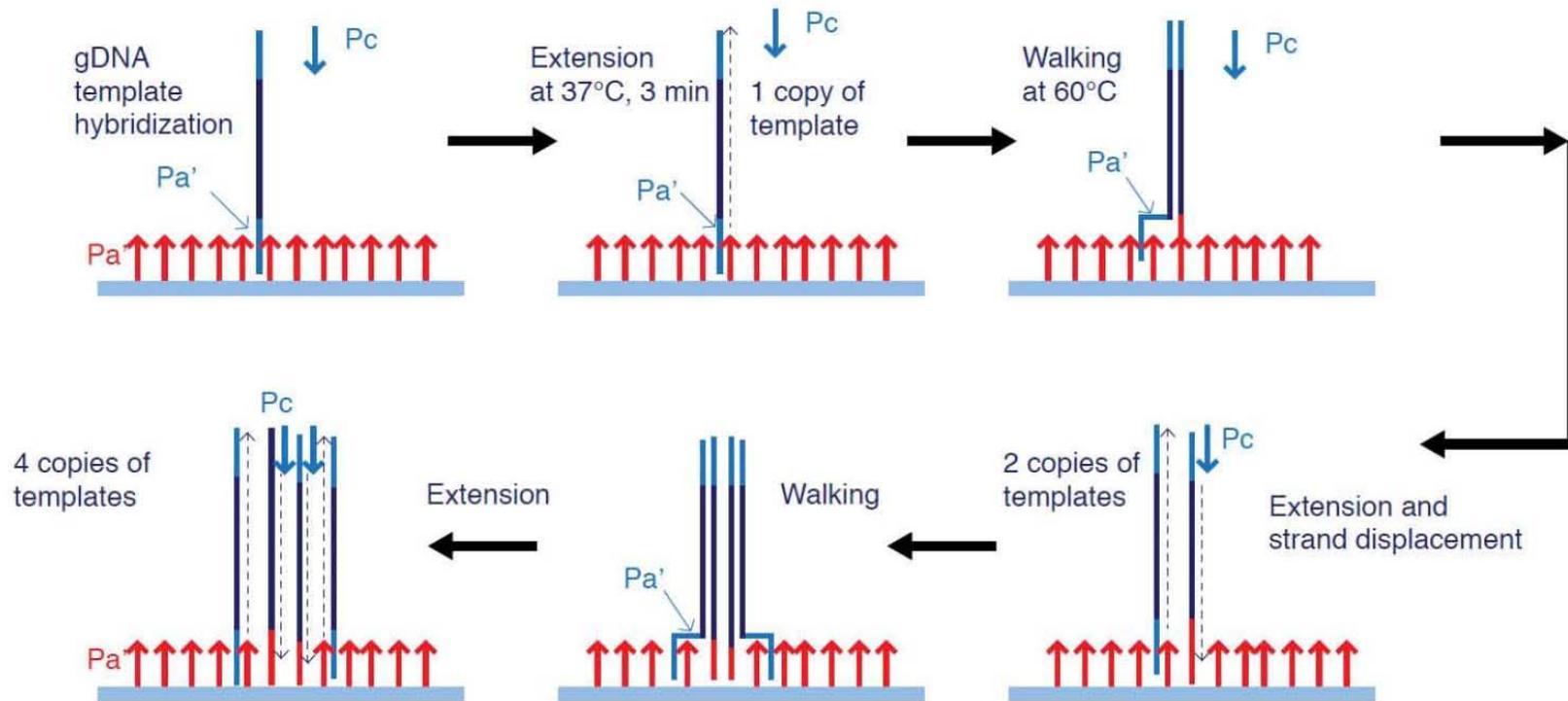
CTAGGTAGCTAGTGG
GCTLIFECISGATAG
C4-LETTERWORDT
GCTATATCGTAGCTG



Next Generation Sequencing : Amplified Single Molecule Sequencing “Polony” PCR

21/132

Wildfire amplification : SOLiD



Wildfire chemistry schematic.



Esquema general

1. Preparación biblioteca
2. 2 opciones
 1. PCR por emulsión
 2. PCR polony en una transparencia
3. 3 opciones
 1. Amplificación de puentes (Illumina)
 2. Amplificación por temperatura (SOLiD)
 3. **Pirosecuenciación (454)**



3.3 Pirosecuenciación

- Usa otra vez microesferas a las que se unen varios fragmentos
- Usando primers se produce la replicación

Next Generation Sequencing

Different platforms

Andy Vierstraete,
Department of Biology,
Ghent University. June 2012

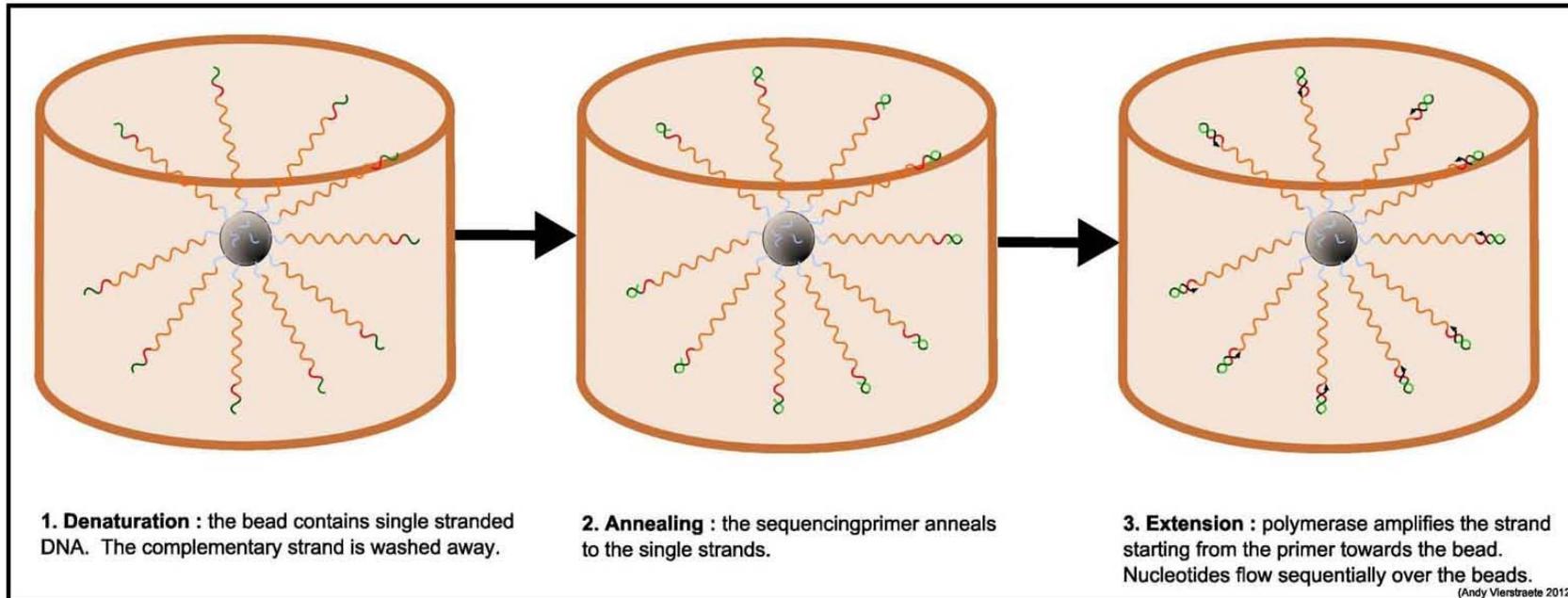
```
CTAGGTAGCTAGTCC  
GCTLIFECISGATAG  
C4-LETTERWORDT  
GCTATATCGTAGCTG
```



26/132

Next Generation Sequencing : Amplified Single Molecule Sequencing

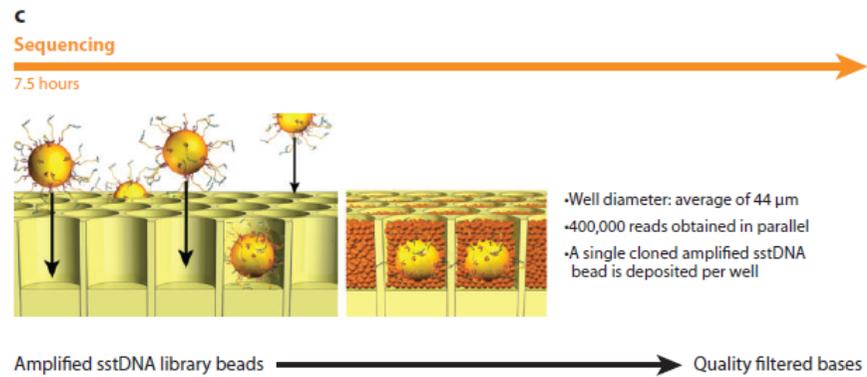
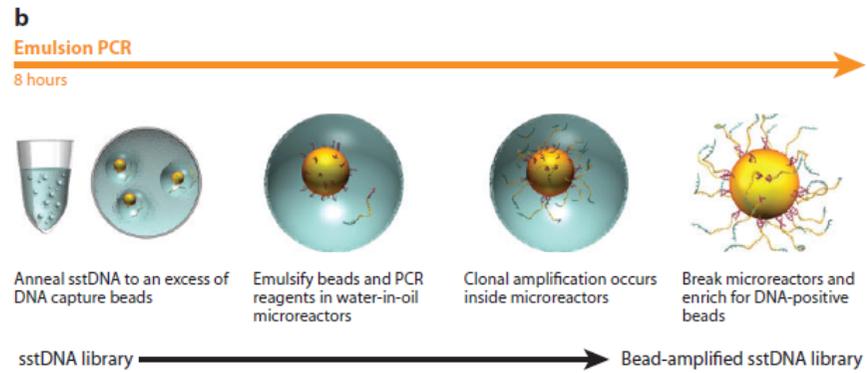
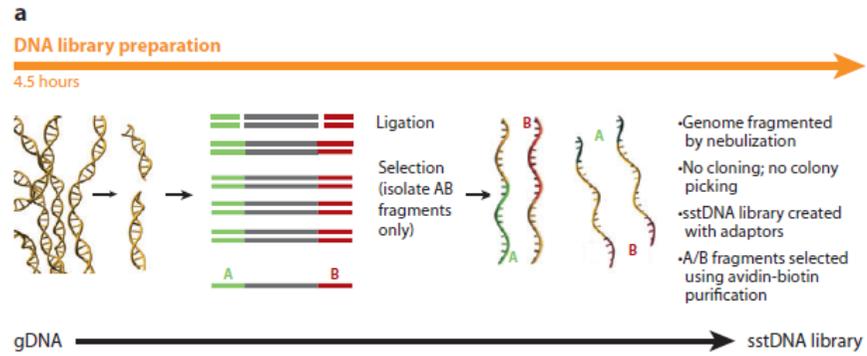
454 Sequencing / Roche Pyrosequencing





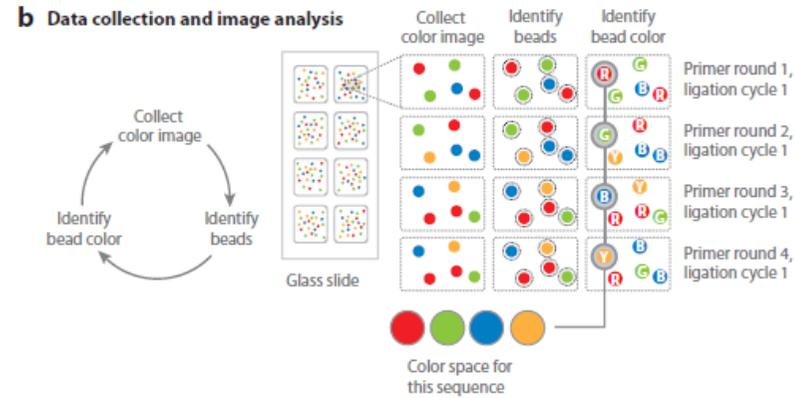
Falta el último paso

- Secuenciación: lectura de los trozos originales y replicados
- Bioquímica algo más complicada

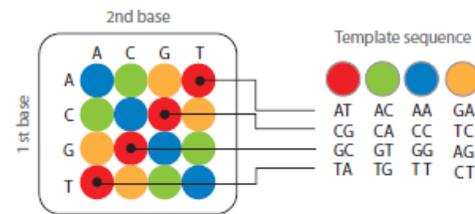


SOLID

codificación por color



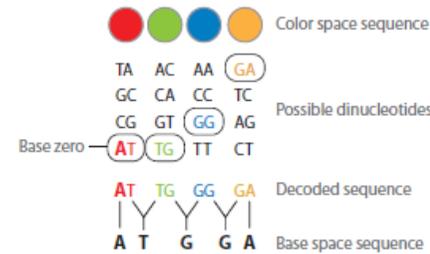
Possible dinucleotides encoded by each color



Double interrogation



Decoding



Comparación de los métodos existentes

	Feature generation	Sequencing by synthesis
454	Emulsion PCR	Polymerase (pyrosequencing)
Solexa	Bridge PCR	Polymerase (reversible terminators)
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)
Polonator	Emulsion PCR	Ligase (nonamers)
HeliScope	Single molecule	Polymerase (asynchronous extensions)

	Cost per megabase	Cost per instrument	Paired ends?	1° error modality	Read-length
454	~\$60	\$500,000	Yes	Indel	250 bp
Solexa	~\$2	\$430,000	Yes	Subst.	36 bp
SOLiD	~\$2	\$591,000	Yes	Subst.	35 bp
Polonator	~\$1	\$155,000	Yes	Subst.	13 bp
HeliScope	~\$1	\$1,350,000	Yes	Del	30 bp



Datos reales - espacio de nucleótidos

- Solexa (Illumina)

@ SRR002051.1: 8: 1: 325: 773 Longitud = 33
AAAGAACATTAAAGCTATATTATAAGCAAAGAT

+ SRR002051.1: 8: 1: 325: 773 Longitud = 33
IIIIIIIIIIIIIIIIIIIIII @ | \$) -

@ SRR002051.2: 8: 1: 409: 432 Longitud = 33
AAGTTATGAAATTGTAATTCCAATATCGTAAGC

+ SRR002051.2: 8: 1: 409: 432 Longitud = 33
IIIIIIIIIIIIIIIIIIIIII07

@ SRR002051.3: 8: 1: 488: 490 Longitud = 33
AATTTCTTACCATATTAGACAAGGCACTATCTT

+ SRR002051.3: 8: 1: 488: 490 Longitud = 33
IIIIIIIIIIIIIIIIIIIIII + i



Datos reales - espacio de color

- datos SOLID

> 1_24_47_F3

T1.1.23..0120230.320033300030030010022.00.0201.0201

> 1_24_52_F3

T2.3.21..2122321.213110332101132321002.11.0111.1222

> 1_24_836_F3

T0.2.22..2222222.010203032021102220200.01.2211.2211

> 1_24_1404_F3

T2.3.30..2013222.222103131323012313233.22.2220.0213

> 1_25_202_F3

T0.3213.111202312203021101111330201000313.121122211

> 1_25_296_F3

T0.1130.100123202213120023121112113212121.013301210



Espacio de color ...

- AA, CC, GG, TT : 0
 - AC, CA, GT, TG : 1
 - AG, CT, GA, TC : 2
 - AT, CG, GC, TA : 3
-
- T2.3.21..212232



Diferencia de salida de datos entre las tres plataformas

- espacio de nucleótidos vs. espacio de color
- Longitud de lecturas cortas
454 (400 ~ 500 pb) > SOLID (70 pb) ~ Solexa (36 ~ 120 pb)



Veremos

- Historia
- Plataformas NGS
- **Tercera generación: Secuenciadores personales**
- Aplicaciones
- Retos bioinformáticos



Tercera generación

- El objetivo es secuenciar una molécula de DNA completa, sin fragmentación
- 3 compañías:
 - **Oxford Nanopore Technology**
 - Pacific Biosciences (optical wavewide)
 - Helico (bancarrota 2015)



Secuenciación por nanoporos

- Recordemos lo que vimos la primera semana sobre la secuenciación de trozos pequeños por electroforesis



electroforesis en gel

- Se trata de separar los fragmentos por tamaño
- Se meten en gel y se aplica un campo eléctrico, la velocidad es inversamente proporcional al tamaño
- Se separan así por longitudes
- Usando trozos de referencia se puede usar para medir la longitud



Secuenciar DNA: chain termination method

- El “chain termination method” se basa en el anterior (gel electrophoresis)
- Tenemos un fragmento de DNA desconocido s, hacemos muchas copias
- Paso 1: conseguir que haya 4 tubos de ensayo A, C, G, T cada uno conteniendo los prefijos de s que terminan en A (C,G,T)



Chain termination method (3)

- Sólo sirve para fragmentos de hasta 1000 bp (más da demasiados errores)
- Puede dar errores de lectura del resultado (llamados errores de secuenciación):
 - Inserción
 - Borrado
 - Sustitución



Secuenciación por nanoporos

- Usa **electroforesis** para transportar una muestra desconocida por un orificio de diámetro 10^{-9} metros
- El sistema contiene una **solución electrolítica** a la que se aplica un campo eléctrico constante



Secuenciación por nanoporos

- La **densidad** de la corriente eléctrica **depende** de las dimensiones del nanoporo y de la **composición del DNA** o RNA que ocupa el nanoporo
- Secuenciar es posible porque, cuando están cerca de los nanoporos, las muestras causan **cambios característicos en la densidad de la corriente**



Secuenciación por nanoporos

- Problemas: resolución, cuesta ver base por base
- Resuelto algorítmicamente????

Gadgets

<https://nanoporetech.com/>

- MinION: secuenciador portátil por \$1000 (sirve para 10-20 Gbp), recambios por \$800

1 genoma humano = 3 Gbp



El futuro: materiales no orgánicos



Which nanopore device is best for you?

[Enlace](#)



Flongle



MinION



GridION X5



PromethION P24/P48

Long read, direct DNA/RNA/epigenetic sequencing, scalable, real time/rapid, on-demand sequencing that is easy to use and install.

- ✓ Your portable device for smaller, individual, rapid tests.
- ✓ When you don't want to multiplex samples or start a larger run.
- ✓ Amplicons, panels/targeted sequencing, quality testing and more.
- ✓ For use with MinIT or a laptop.

- ✓ Your personal sequencer, putting you in control.
- ✓ Whether in your lab or out in the field.
- ✓ Whole genomes/exomes, metagenomics, targeted sequencing, whole transcriptome (cDNA), smaller transcriptomes (direct RNA), multiplexing for smaller samples and more.
- ✓ For use with MinIT or a laptop.

- ✓ High throughput sequencing, in modular form (up to 5 flow cells) to be on-demand.
- ✓ For your lab or to offer as a service.
- ✓ Larger genomes or projects, whole transcriptomes (direct RNA or cDNA) or where you have larger numbers of samples and more.
- ✓ Compute included for real time data analysis and easy installation.

- ✓ Very high throughput sequencing, in modular form (up to 48 flow cells) to be on-demand.
- ✓ For your lab or as a service.
- ✓ Larger genomes or projects, whole transcriptomes (direct RNA or cDNA), very large numbers of samples and more.
- ✓ Compute included for real time data analysis and easy installation.



Veremos

- Historia
- Plataformas NGS
- Secuenciadores personales
- **Aplicaciones**
- Retos bioinformáticos

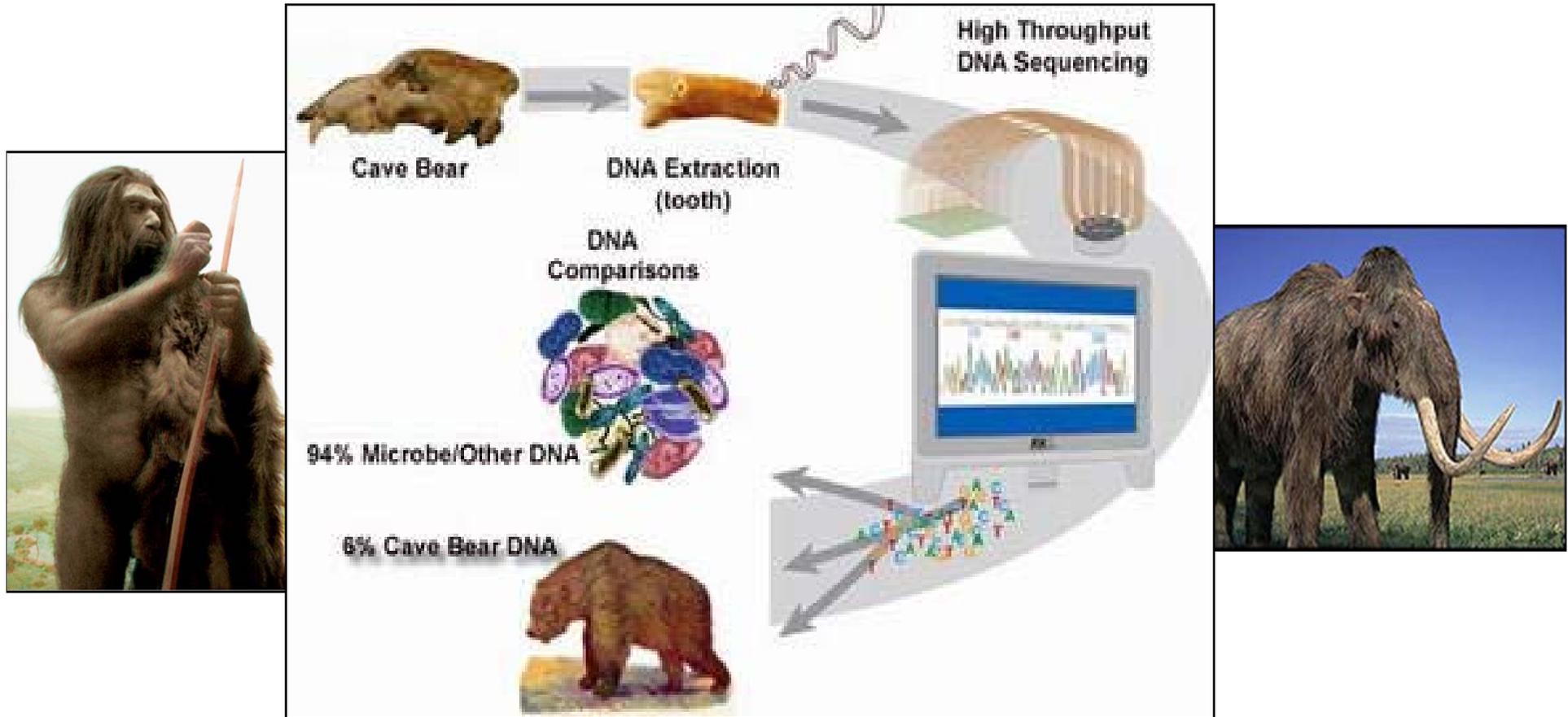


Aplicaciones

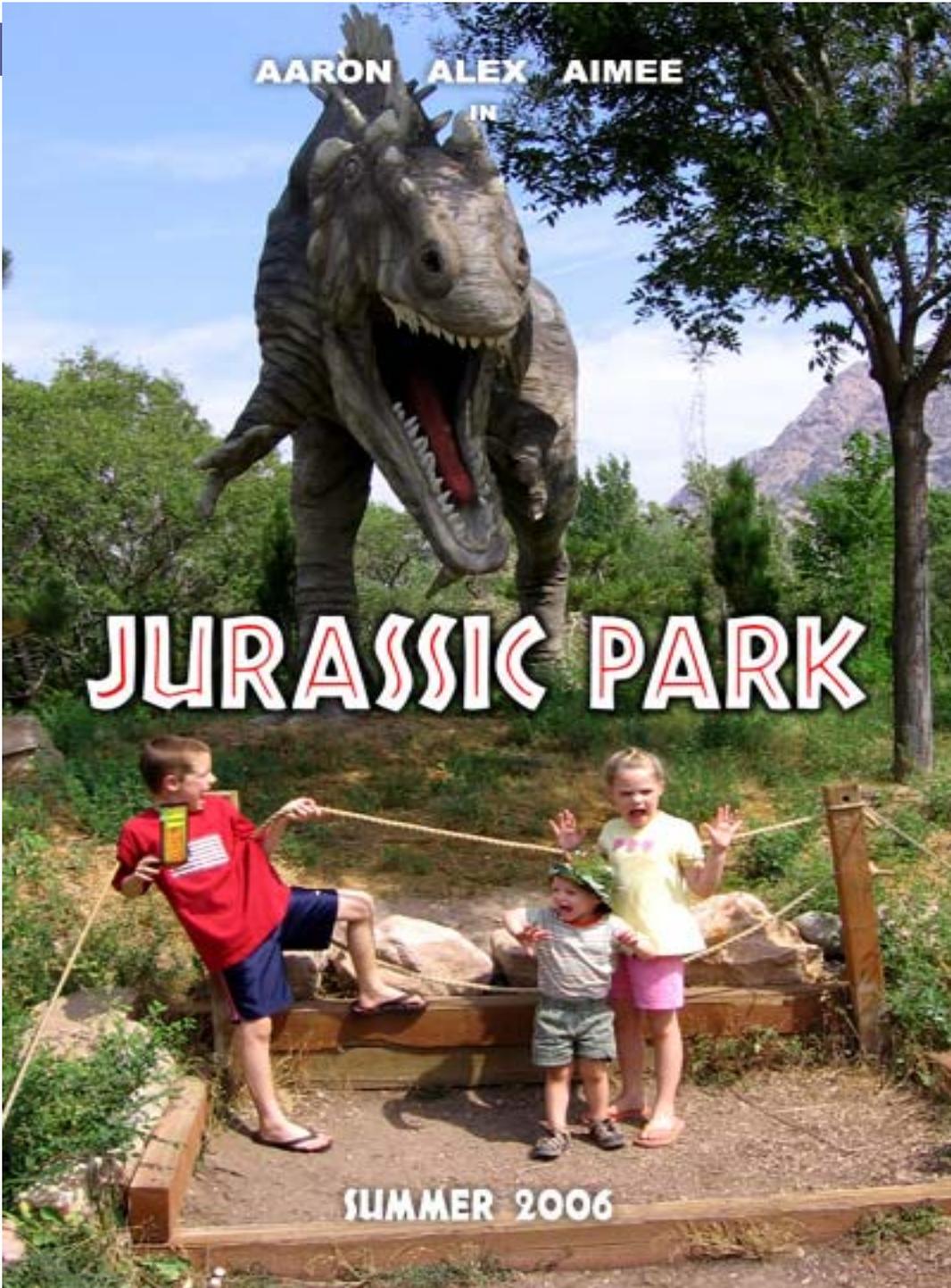
- Ensamblaje de novo de genoma
- Resecuenciación de genoma
- RNA-Seq (la expresión génica, estructura exón-intrón, los pequeños perfiles de RNA, y las mutaciones)
- CHIP-Seq (interacción proteína-DNA)
- perfiles epigenéticos

Los genomas antiguos resucitados

- Estado de la muestra degradado → secuenciación mtDNA
- genomas nucleares de restos antiguos: oso de las cavernas, Mammuth, Neanderthal (10^6 bp)

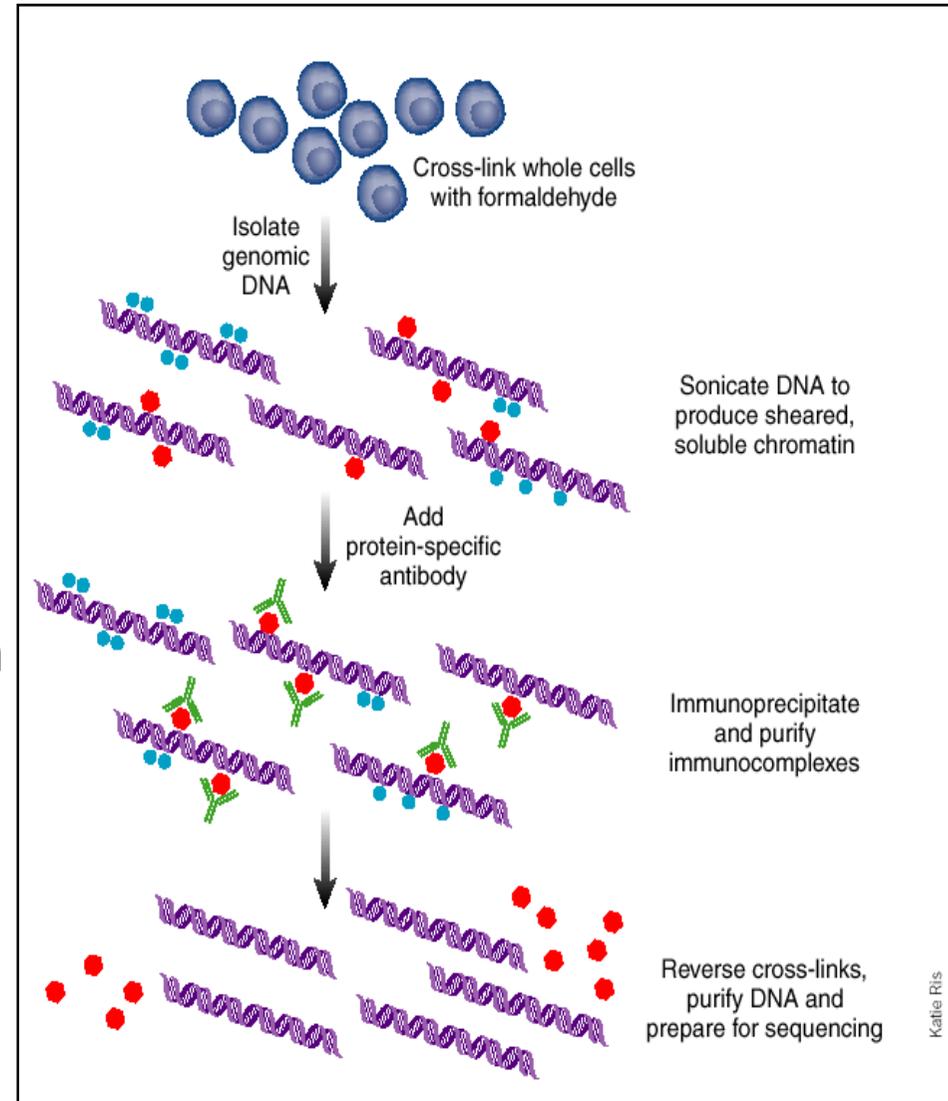


Los problemas de contaminación: los humanos modernos y coisolación del DNA bacteriano



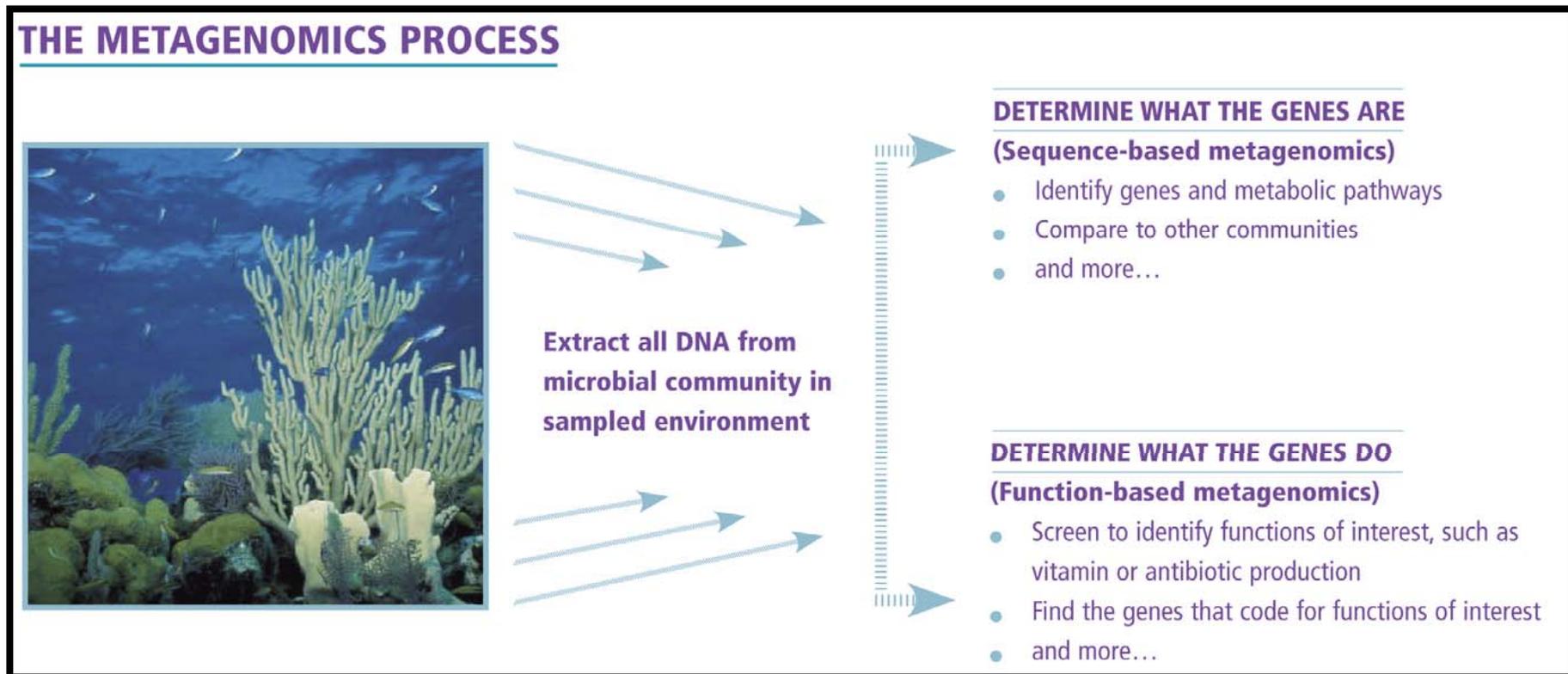
Epigenética: interacciones proteína-ADN a través de la secuenciación

- pieza clave en la regulación de la expresión génica
- Recientemente, los estudios en todo el genoma de interacciones proteína-ADN
- Analizar factor de transcripción / estados de histona en el genoma humano
- Mejorar nuestra comprensión de la expresión génica en el contexto de los estímulos ambientales específicos



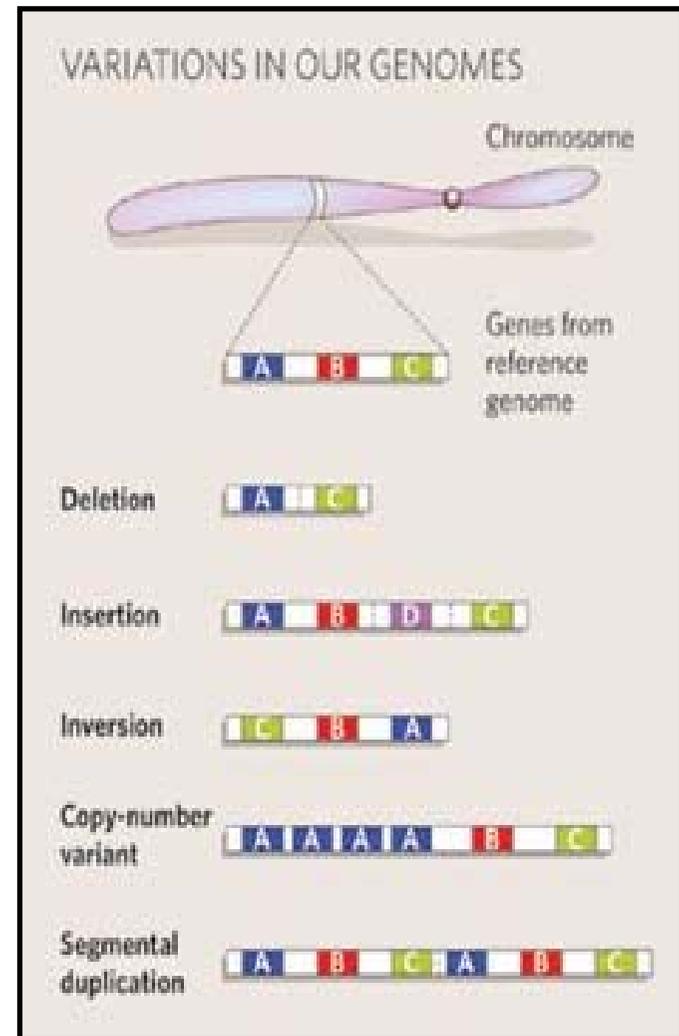
La metagenómica

- Caracterización de la biodiversidad que se encuentra en la Tierra
- El creciente número de genomas secuenciados permite interpretar secuencias parciales obtenidas por muestreo directo de nichos ambientales.
- Ejemplos: lecho marino, el suelo, los arrecifes de coral, microbioma humano que puede variar según el estado de salud del individuo



Definición de variabilidad en muchos genomas humanos

- variantes comunes aún no han explicado completamente la genética de enfermedades complejas → alelos raros también contribuyen
- También variantes estructurales, inserciones y deleciones grandes y pequeños
- Acelerar la investigación biomédica





Veremos

- Historia
- Plataformas NGS
- Secuenciadores personales
- Aplicaciones
- **Retos bioinformáticos**



Cantidad enorme de datos

- Operación típica con SOLiD:

- ~ Archivo de imagen 2T

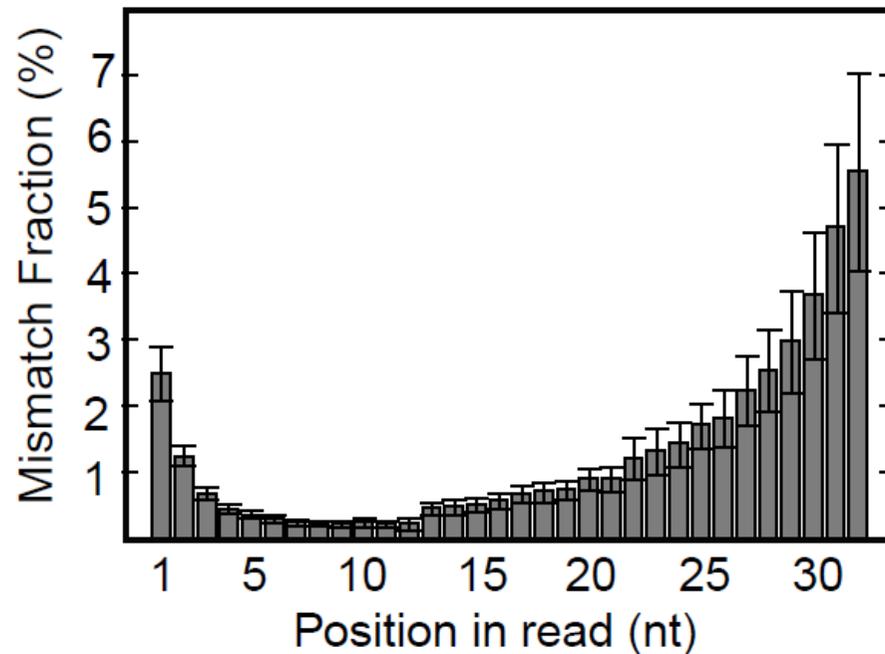
- ~ archivo de texto 120G para extracción de conocimiento

- ~ 75 M fragmentos cortos por muestra

Métodos eficientes para el almacenamiento y gestión de datos

Requiere cálculo paralelo de alto rendimiento

Tasas de error de secuenciación considerables



análisis de imágenes de alta calidad para
identificación de bases



Retos bioinformáticos

- Métodos eficientes para almacenar, recuperar y procesar gran cantidad de datos
- Reducir los errores en el análisis de imágenes e identificación de bases (“base calling”)
- Métodos rápidos y precisos para alineación y montaje del genoma
- Nuevos algoritmos de extracción de conocimiento